Vector Calculus

Central to vector calculus is the concept of a function f, which is a quantity that relates 2 quantities, the inputs $x \in \mathbb{R}^D$ (the *domain*) and targets $f(x)$ (the *image/codomain*), to each other.

$$\begin{aligned} f : \mathbb{R}^D &\to \mathbb{R} \\ x &\mapsto f(x) \end{aligned}$$

A function f assigns every input x exactly one function value $f(x)$.

> **Example**
>
> Use the dot product as a special case of an inner product. The function $f(x) = x^T x, x \in \mathbb{R}^2$ would be
>
> $$\begin{aligned} f : \mathbb{R}^2 &\to \mathbb{R} \\ x &\mapsto x_1^2 + x_2^2 \end{aligned}$$

Differentiation of Univariate Functions

Derivative. For $h > 0$ the derivative of f at x is defined as the limit

$$\frac{df}{dx} := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

The derivative of f points in the direction of steepest ascent of f.

> Derivative of Polynomial

We want to compute the derivative of $f(x) = x^n, n \in \mathbb{N}$. We may already know that the answer will be nx^{n-1}, but we want to derive this result using the definition of the derivative as the limit of the difference quotient.

Using the definition of the derivative in (5.4), we obtain

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \tag{5.5a}$$

$$= \lim_{h\to 0} \frac{(x+h)^n - x^n}{h} \tag{5.5b}$$

$$= \lim_{h\to 0} \frac{\sum_{i=0}^{n} \binom{n}{i} x^{n-i}h^i - x^n}{h}\,. \tag{5.5c}$$

We see that $x^n = \binom{n}{0}x^{n-0}h^0$. By starting the sum at 1, the x^n-term cancels, and we obtain

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \lim_{h\to 0} \frac{\sum_{i=1}^{n} \binom{n}{i} x^{n-i}h^i}{h} \tag{5.6a}$$

$$= \lim_{h\to 0} \sum_{i=1}^{n} \binom{n}{i} x^{n-i}h^{i-1} \tag{5.6b}$$

$$= \lim_{h\to 0} \binom{n}{1} x^{n-1} + \underbrace{\sum_{i=2}^{n} \binom{n}{i} x^{n-i}h^{i-1}}_{\to 0 \text{ as } h\to 0} \tag{5.6c}$$

$$= \frac{n!}{1!(n-1)!}x^{n-1} = nx^{n-1}\,. \tag{5.6d}$$

Taylor Series

The Taylor series is a representation of a function f as an infinite sum of terms. These terms are determined using derivatives of f evaluated at x_0.

Taylor Polynomial. The Taylor polynomial of degree n of $f : \mathbb{R} \to \mathbb{R}$ at x_0 is defined as:

$$T_n(x) := \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

where $f^{(k)}(x_0)$ is the k-th derivative of f at x_0 and $\frac{f^{(k)}(x_0)}{k!}$ are the coefficients of the polynomial.

> For a smooth function $f \in C^\infty$, $f : \mathbb{R} \to \mathbb{R}$, the Taylor series of f at x_0 is defined as
>
> $$T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$
>
> For $x_0 = 0$, we obtain the *Maclaurin series* as a special instance of the Taylor series.
>
> If $f(x) = T_\infty(x)$, then f is called *analytic*.

Differentiation rules

Product rule:

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$$

Quotient rule

$$(\frac{f(x)}{g(x)})' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$$

Sum rule

$$(f(x) + g(x))' = f'(x) + g'(x)$$

Chain Rule

$$(g(f(x)))' = (g \circ f)'(x) = g'(f(x))f'(x)$$

where $g \circ f$ denotes function composition $x \mapsto f(x) \mapsto g(f(x))$

Partial Differentiation and Gradients

The generalisation of the derivative to functions of several variables is the *gradient*.

Find the gradient to the function f with respect to x by varying one variable at time and keeping the others constant. The gradient is then the collection of these partial derivatives.

Partial Derivative. For a function $f : \mathbb{R}^m \to \mathbb{R}, x \mapsto f(x), x \in \mathbb{R}^n$ of n variables $x_1, \ldots, x_n$ we define the partial derivatives as:

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(x)}{h} \\ &\vdots \\ \frac{\partial f}{\partial x_n} &= \lim_{h \to 0} \frac{f(x_1, \ldots, x_{n-1}, x_n + h) - f(x)}{h} \end{aligned}$$

and collect them in the row vector, the row vector is also called *Jacobian,*

$$\nabla_x f = \mathrm{grad} f = \frac{df}{dx} = [\frac{\partial f(x)}{\partial x_1} \frac{\partial f(x)}{\partial x_2} \cdots \frac{\partial f(x)}{\partial x_n}] \in \mathbb{R}^{1 \times n}$$

where n is the number of variables and 1 is the dimension of the image of f.

For any vector v tangent to the level surface, the gradient is perpendicular to it, $\nabla f \cdot v = 0$.

The ∇f is normal vector to the tangent plane.

Basic rules of partial differentiation

Pay attention here the gradients involve vectors and matrices, and matrix multiplication is not commutative, the order is important.

<u>Product rule</u>

$$\frac{\partial}{\partial x}(f(x)g(x)) = \frac{\partial f}{\partial x} g(x) + f(x) \frac{\partial g}{\partial x}$$

<u>Sum rule</u>

$$\frac{\partial}{\partial x}(f(x) + g(x)) = \frac{\partial f}{\partial x} + \frac{\partial g}{\partial x}$$

<u>Chain rule</u>

$$\frac{\partial}{\partial x}(g \circ f)(x) = \frac{\partial}{\partial x}(g(f(x))) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial x}$$

Chain rule

Consider a function $f : \mathbb{R}^2 \to \mathbb{R}$ of 2 variables x_1, x_2. And $x_1(t), x_2(t)$ are themselves functions of t. To compute the gradient of f with respect to t:

$$\frac{df}{dt} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}$$

If $f(x_1, x_2)$ is a function of x_1 and x_2, where $x_1(s,t)$ and $x_2(s,t)$ are functions to s, t, the chain rule yields

$$\begin{aligned} \frac{\partial f}{\partial s} &= \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s} \\ \frac{\partial f}{\partial t} &= \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t} \end{aligned}$$

and the gradient is obtained by the matrix multiplication

$$\frac{\mathrm{d}f}{\mathrm{d}(s,t)} = \frac{\partial f}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial(s,t)} = \underbrace{\begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix}}_{= \frac{\partial f}{\partial \boldsymbol{x}}} \underbrace{\begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix}}_{= \frac{\partial \boldsymbol{x}}{\partial(s,t)}} .$$

The Hessian Matrix

The notations for higher-order gradients,

- $\frac{\partial^n f}{\partial x^n}$ is the n-th partial derivative of f to x
- $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)$ is the partial derivative obtained to x then to y

The Hessian is the collection of all second-order partial derivatives.

If $f(x, y)$ is a twice (continuously) differentiable function, then

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$$

the order of differentiation doesn't matter, the corresponding *Hessian matrix*

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$$

is symmetric, and the Hessian is denoted as $\nabla^2_{x,y} f(x, y)$.

> For a square matrix A,
>
> - PD: $x^T Ax > 0$, ND: $x^T Ax < 0$
> - PSD: $x^T Ax \geq 0$, NSD: $x^T Ax \leq 0$

- If Hessian is Positive Defined at a point, the function is locally convex. Its critical point is local minimum.
- If Hessian is Negative Defined, then the function is locally concave. Its critical point is local maximum.

The Hessian measures the *local curvature* at some point (x, y), and the gradient tells the *local slope*.

Gradients of Vector-Valued Functions

For a function $f : \mathbb{R}^n \to \mathbb{R}^m$ and a vector $x = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$, the corresponding vector of function values is

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix} \in \mathbb{R}^m$$

The gradient of $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ to $\boldsymbol{x} \in \mathbb{R}^n$ by collecting these partial derivatives:

$$\frac{\mathrm{d}\boldsymbol{f}(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} = \left[\boxed{\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1}} \cdots \boxed{\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n}}\right] = \left[\boxed{\begin{matrix}\frac{\partial f_1(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1}\end{matrix}} \begin{matrix}\cdots \\ \\ \cdots\end{matrix} \boxed{\begin{matrix}\frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_n}\end{matrix}}\right] \in \mathbb{R}^{m \times n}.$$

Jacobian. The collection of all first-order partial derivatives of a vector-valued function $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ is called the *Jacobian.* The Jacobian $\boldsymbol{J}$ is an $m \times m$ matrix,

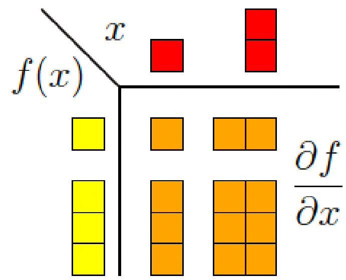
$$\boldsymbol{J} = \nabla_{\boldsymbol{x}} \boldsymbol{f} = \frac{\mathrm{d}\boldsymbol{f}(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix}\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n}\end{bmatrix}$$

$$= \begin{bmatrix}\frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n}\end{bmatrix},$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad J(i,j) = \frac{\partial f_i}{\partial x_j}.$$

A summary of the dimensions of those derivatives.

- $f : \mathbb{R} \to \mathbb{R}$, the gradient is a scalar
- $f : \mathbb{R}^D \to \mathbb{R}$, the gradient is a $1 \times D$ row vector
- $f : \mathbb{R} \to \mathbb{R}^E$, the gradient is an $E \times 1$ column vector
- $f : \mathbb{R}^D \to \mathbb{R}^E$, the gradient is an $E \times D$ matrix



Example

We are given

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}\,, \qquad \boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^M, \quad \boldsymbol{A} \in \mathbb{R}^{M\times N}, \quad \boldsymbol{x} \in \mathbb{R}^N\,.$$

To compute the gradient $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{x}$ we first determine the dimension of $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{x}$: Since $\boldsymbol{f} : \mathbb{R}^N \to \mathbb{R}^M$, it follows that $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{x} \in \mathbb{R}^{M\times N}$. Second, to compute the gradient we determine the partial derivatives of f with respect to every x_j:

$$f_i(\boldsymbol{x}) = \sum_{j=1}^{N} A_{ij}x_j \implies \frac{\partial f_i}{\partial x_j} = A_{ij} \tag{5.67}$$

We collect the partial derivatives in the Jacobian and obtain the gradient

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & & \vdots \\ A_{M1} & \cdots & A_{MN} \end{bmatrix} = \boldsymbol{A} \in \mathbb{R}^{M\times N}\,. \tag{5.68}$$

Chain Rule

Consider the function $h : \mathbb{R} \to \mathbb{R}$, $h(t) = (f \circ g)(t)$ with

$$f : \mathbb{R}^2 \to \mathbb{R} \tag{5.69}$$

$$g : \mathbb{R} \to \mathbb{R}^2 \tag{5.70}$$

$$f(\boldsymbol{x}) = \exp(x_1 x_2^2)\,, \tag{5.71}$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = g(t) = \begin{bmatrix} t\cos t \\ t\sin t \end{bmatrix} \tag{5.72}$$

and compute the gradient of h with respect to t. Since $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}^2$ we note that

$$\frac{\partial f}{\partial \boldsymbol{x}} \in \mathbb{R}^{1\times 2}\,, \quad \frac{\partial g}{\partial t} \in \mathbb{R}^{2\times 1}\,. \tag{5.73}$$

The desired gradient is computed by applying the chain rule:

$$\frac{\mathrm{d}h}{\mathrm{d}t} = \frac{\partial f}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{x}}{\partial t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix} \tag{5.74a}$$

$$= \begin{bmatrix} \exp(x_1 x_2^2) x_2^2 & 2\exp(x_1 x_2^2) x_1 x_2 \end{bmatrix} \begin{bmatrix} \cos t - t\sin t \\ \sin t + t\cos t \end{bmatrix} \tag{5.74b}$$

$$= \exp(x_1 x_2^2)\big(x_2^2(\cos t - t\sin t) + 2x_1 x_2(\sin t + t\cos t)\big)\,, \tag{5.74c}$$

where $x_1 = t\cos t$ and $x_2 = t\sin t$; see (5.72).

Useful identities for computing gradients

$$\frac{\partial}{\partial \boldsymbol{X}} \boldsymbol{f}(\boldsymbol{X})^\top = \left(\frac{\partial \boldsymbol{f}(\boldsymbol{X})}{\partial \boldsymbol{X}}\right)^\top \tag{5.99}$$

$$\frac{\partial}{\partial \boldsymbol{X}} \operatorname{tr}(\boldsymbol{f}(\boldsymbol{X})) = \operatorname{tr}\left(\frac{\partial \boldsymbol{f}(\boldsymbol{X})}{\partial \boldsymbol{X}}\right) \tag{5.100}$$

$$\frac{\partial}{\partial \boldsymbol{X}} \det(\boldsymbol{f}(\boldsymbol{X})) = \det(\boldsymbol{f}(\boldsymbol{X})) \operatorname{tr}\left(\boldsymbol{f}(\boldsymbol{X})^{-1} \frac{\partial \boldsymbol{f}(\boldsymbol{X})}{\partial \boldsymbol{X}}\right) \tag{5.101}$$

$$\frac{\partial}{\partial \boldsymbol{X}} \boldsymbol{f}(\boldsymbol{X})^{-1} = -\boldsymbol{f}(\boldsymbol{X})^{-1} \frac{\partial \boldsymbol{f}(\boldsymbol{X})}{\partial \boldsymbol{X}} \boldsymbol{f}(\boldsymbol{X})^{-1} \tag{5.102}$$

$$\frac{\partial \boldsymbol{a}^\top \boldsymbol{X}^{-1} \boldsymbol{b}}{\partial \boldsymbol{X}} = -(\boldsymbol{X}^{-1})^\top \boldsymbol{a} \boldsymbol{b}^\top (\boldsymbol{X}^{-1})^\top \tag{5.103}$$

$$\frac{\partial \boldsymbol{x}^\top \boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}^\top \tag{5.104}$$

$$\frac{\partial \boldsymbol{a}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}^\top \tag{5.105}$$

$$\frac{\partial \boldsymbol{a}^\top \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a} \boldsymbol{b}^\top \tag{5.106}$$

$$\frac{\partial \boldsymbol{x}^\top \boldsymbol{B} \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{x}^\top (\boldsymbol{B} + \boldsymbol{B}^\top) \tag{5.107}$$

$$\frac{\partial}{\partial \boldsymbol{s}} (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^\top \boldsymbol{W} (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s}) = -2(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^\top \boldsymbol{W} \boldsymbol{A} \quad \text{for symmetric } \boldsymbol{W} \tag{5.108}$$